A Scene Division Method Based on Theme

Fuping Yang[1], Zhichun Yuan[1] and Xi Cheng[2]
[1]*Department of Computer Science and Technology, Chongqing University of Posts and Telecommunications NanShan Street, Chongqing, China*

Keywords: Text-to-scene, Scene division, Multi-scene text, Scene features, Scene themes.

Abstract: The text-to-scene conversion is such a process that converts the input text to 3D scenes automatically based on the natural language processing. The scene division is one of the basic research contents of the text-to-scene conversion, and it mainly identifies and divides the number and structure of scenes in multi-scene text. In this paper, the general expression of the scene is obtained through the analysis of scene features, and then the method for divide multi-scene text which to use scene themes has been proposed. In the experiment, the LDA model is used to divide the multi-scene text, and this method is applied to the standard data set. The experiments show that the expected results are achieved.

1 INTRODUCTION

Language is an important communication tool for human beings, and it can convey people's thoughts and display visual scenes that are imagined in the brain. The concept of "text-to-scene" was first proposed by Coyne and Sproat (Coyne, 2001), which refers to the process of automatically converting a natural language text description into a three-dimensional static scene or animation. Manual building a 3D scene is a time-consuming and intensive process that requires users to adopt and learn professional graphics tools and interfaces, and such requirements limit the number of potential users (Seversky, 2006). The purpose of text-to-scene conversion is to enable people for the non-graphics professional field to realize the scenes in their minds through simple text descriptions, to make communication between each other faster and more intuitive, reduce the process of manually constructing scenes, and improve work efficiency.

At present, the research of the text-to-scene conversion is in its infancy and has not yet formed a complete theoretical system. Some scholars have made a preliminary exploration of the text-to-scene conversion, and there are already a lot of prototype systems. Story Picturing Engine (Joshi, 2004) refers to the process of illustrating a story with suitable pictures. NALIG (Giovanni, 1984) is one of the early projects on generating static 2D scenes from natural language descriptions. Put (Clay, 1996) is a rule-based spatial placement system jointly researched by Silicon and the University of California, which generates an object's placement scene through restricted natural language input. WordsEye (Coyne, 2001) is such a system for automatically converting text into 3D scenes. It relies on a large database of 3D models and poses to depict entities and actions. Compared to previous systems, the system developed at Stanford University (Chang, 2014a 2014b) can infers the implicit relations and partially supports interactive scene manipulation and active learning. SHRLDU (Winograd, 1971) is one of the pioneer systems developed by the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology in the natural language to direct robot movements. It can realize the spatial placement of simple geometry, which is also a significant research result in the field of natural language processing and artificial intelligence. Carsim (Johansson, 2004) is a domain-specific system developed by Lund University that generates simple car accident animations based on a series of Swedish accident reports, victim narratives and official records of officials collected from news articles. Combining the above systems, it can be finded that they all generated the scenes without considering the division of the scene. However, the text language usually contains multiple scenes. It is necessary to divide the scene in the multi-scene text when generating the scenes. In order to resolve the

Yang, F., Yuan, Z. and Cheng, X.
A Scene Division Method Based on Theme.
DOI: 10.5220/0008098000440048
In *Proceedings of the International Conference on Advances in Computer Technology, Information Science and Communications (CTISC 2019)*, pages 44-48
ISBN: 978-989-758-357-5

problem of dividing the multi-scene text, this paper made a preliminary attempt.

The paper is organized as follows. The first section introduces the research background, purpose and significance of text-to-scene conversion. The second section is an overview of the relevant foundations of text-to-scene conversion. The third section is the feature analysis of the scene in the text-to-scene conversion, which contains the concept of the scene and its constituent elements. The fourth section is an introduction of the scene division method based on theme, which includes algorithms, data sets, and analysis of experimental results. The fifth section is the summary of this paper.

2 TEXT-TO-SCENE

The text-to-scene conversion mainly involves scene information extraction, establishment of mapping relationship for text to scene and the generation of 3D scenes. First, natural language processing methods are used to mine and extract the scene information contained in the text. The scene information includes the scene objects, the visual attributes of the object, and the spatial relationship between the objects, etc. Secondly, the extracted scene information is associated with the model library, which is a transition phase from computer linguistics to computer graphics. Finally, scene generation involves the production and presentation of 3D scenes or animations. The basic theory of existing text-to-scene conversion reasearch is not sufficient, and the related papers mainly analyze the various modules that constitute the text-to-scene conversion system. Most of those research works focus on constructing a standard expression of text and mapping the expression to the relationship between the three-dimensional world coordinates of the graphics. This mapping reflects the entity noun to the model and the layout information to the layout of models. With the development of text-to-scene conversion, people began to foucs on the modeling of “scene”. Usually, a scene is complicated. The natural language description of the scene is simple and abstract, not enough to reflect every aspect of the scene, which leads to the generated scene sometimes does not satisfy people's common sense of the real world. From the perspective of “scene”, it is significant to discuss the composition of the scene.

3 SCENE FEATURE ANALYSIS

3.1 Scene Definition

A scene refers to a series of objects combination associated with theme in a specific time and space environment. A scene consists of a five-tuple, mainly including scene theme, scene time, scene space, scene objects, and scene relationships.

3.1.1 Scene Theme

The scene comes from the description of the text in text-to-scene conversion. Scene theme is a general expression of the connotation of the scene. In fact, every text has a theme. The scene also should be the same as the text, with the corresponding scene theme. Most of the times, the scene theme is the same as the text theme, but it is not all. For example, “John is eating breakfast”, it can be finded that the theme of the scene and the theme of the text are to eat breakfast. There is usually no scene in the text describing politics. Thus, there is no scene theme. The scene theme is the core element of the scene, and other elements are related to the scene theme. The objects in the scene can be determined through the scene theme. For example, the theme of the scene is birthday, and there are cake and candles in the scene, but there is no sea. Scene space and scene time can also be inferred by scene theme. For example, the theme of the scene is watching TV and scene space may be in the living room. The theme of the scene is snowing and it can be known that the scene time is in winter.

3.1.2 Scene Time

Scene time refers to the time when the scene occurred or appeared. Each Scene has a definite scene time. The scene time of a static scene can be a constant value, a positive interval, or expressed as infinity. The scene time of the dynamic scene is a positive interval. A static scene refers to a scene in which the state of all objects does not change. For example, the scene describing the campus layout is a static scene. The scene time describing the campus layout may be a certain day or a certain moment. However, scene time does not generally appear in scene texts that describe static scenes. A dynamic scene refers to a scene in which at least one object's state is changed. For example, the scene describing the basketball game is a dynamic scene. The scene time describing the basketball game may be a certain time period.

3.1.3 Scene Space

Scene space refers to the place where the scene exists. Each Scene has a scene space. Scene space can be divided into natural or artificial, indoor or outdoor, etc. The scene space can be a country, a city, or a particular location. For instance, we regard the Great Wall as a scene, and then the scene space is China, more specifically Beijing.

3.1.4 Scene Objects

Scene objects are things that can be visualized in the scene. The scene objects usually refer to entities that exist in the scene, and entity refers to things that can be directly visualized in the scene. Scene objects are crucial in the scene, and it is one of the most basic factors that make up the scene. A scene without a scene object is like a blank drawing paper. From the perspective of part of speech, scene objects belong to the category of nouns, but not all nouns represent a scene object. The apple, table, wind, China, etc. are called scene objects, scene objects are things that exist in the world or can be represented by existing things. The apple and table represent a kind of existence and concrete things in the real world, so they also are called entities. For entities, they have their own visual attributes. The visual attributes of the entity include color, size, shape, texture, etc. In addition, there are some nouns that reflect the process or result of humans' cognitive in the objective world. They are not referring to the existence and specific things, but the invisible concepts. They are not suitable for visualization. For example, spiritual, material, and friendship are typical of these words.

3.1.5 Scene Relationships

Scene relationship includes the spatial and non-spatial relationship between objects in the scene. The spatial relations are a crucial component in descriptions of scenes. Spatial relations define the basic layout information of scenes, which includes the orientation relationship, the distance relationship, and the topological relationship. For example, "the computer is on the table", "John is 5 feet in front of the tree", and "a picture is hanging on the wall". Spatial relations are often denoted by prepositions such as on, under, beyond, in front of, etc. In the description of the spatial relationship, the specific location information of the object is also related to the size and shape of the object. Non-spatial relationships are not obviously reflected in the scene, but they are also important in the scene. The non-spatial relationships between the objects have the part-of relation, the container-of relation, etc. For example, "arm of the chair" is the part-of relation. "Bowl of cherries" is the container-of relation.

4 SCENE DIVISION

An article consists of a series of words, and the most important words in multi-scene text are the entity noun. Therefore, in order to divide the multi-scene text, you must first annotate the entity noun with topic IDs. It is not based on words, but on topic IDs assigned by LDA model. This increases sparsity because the word space is reduced to a lower dimensional topic space. The topic model must be trained on a document similar to the content of the test document to make the method effective. A sentence is assumed to be the smallest basic unit. We introduced a window parameter that defines the number of sentences contained in a window, and then the similarity of two adjacent windows is calculated. We cannot declare the value of window parameter in advance, and this is conditioned on the multi-scene text that is divided. To calculate the similarity, we exclusively use the topic IDs assigned to the entity noun. Assuming an LDA model with T topics, the frequency of each topic in a window block is counted to compose into a T-dimensional vector. The cosine similarity between vectors is calculated. If the value is close to zero, it denotes that the two window block edges are not associated. They belong to two different scenes. The value is close to one, it indicates that the two window block margins are associated. They belong to a scene. In this experiment, we identify the scene boundaries by setting a threshold.

4.1 LDA Model

LDA is a generative model for text and other collections of discrete data introduced by Blei et al (Blei, 2003), which is a classic probabilistic topic model. LDA model contains three layers of words, topics, and documents. Blei added the Bayesian prior probability to the PLSA algorithm, which gave birth to the LDA algorithm. LDA is an unsupervised machine learning model that can be used to identify potential topic information in large-scale document sets or corpora. LDA has wide application in the fields of natural language processing and document classification. This method assumes that each word in a document is extracted from the topic. Model training evaluates two distributions: document-topic distribution, topic-word distribution. Since LDA is a

probability generation model, for each document in the corpus, the following generative process is defined. For each document, a topic is extracted from the topic distribution. Then a word is extracted from the word distribution corresponding to the extracted topic. Repeat the above process until each word in the document is traversed.

4.2 Data Sets

The performance of the introduced method is demonstrated using data sets from the primary school compositions, which contains of 100 compositions depicting different scenes. The number of sentences in each composition is 5 to 20 and they are all made up of one theme.

4.3 Evaluation

The performance of the approach based on theme is evaluated by accuracy, recall, and F1 values. In addition to the method assessment, you also need to specify the following parameters for the LDA model: The number of topics T; the value of alpha, hyper-parameter of LDA; the value of beta, also the hyper-parameter of LDA; the number of Gibbs sampling iterations. In this paper, the Bayesian statistical standard method is used to analyse the fitting of the model to the corpus by setting the number of different scene topics, and finally determine the number of optimal scene topics. Figure 1 illustrates the experimental results of setting different topic number. As shown in Figure 1, when the number of topics T is 2, the value of log-likelihood is the largest, and then begins to gradually decrease. At this time, the model fits the corpus best. The values of the remaining parameters in the experiment are as follows: The value of alpha is 50/T; the value of beta is 0.5; the number of Gibbs sampling iterations is 2000.

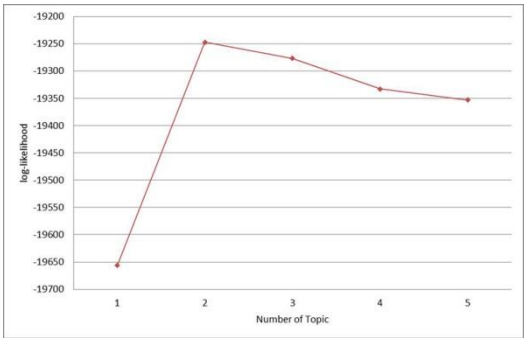

Figure 1: Number of Topic and log-likelihood curve.

The compositions from different topics are combined to form a multi-scene text as a test set, the number of sentences is 15 to 45 for each multi-scene text. We extract the nouns in the sentence through word segmentation and part-of-speech tagging, and then manually choose the entity nouns in them. In the experiment, we set up different windows for comparison. For the same window, the accuracy, recall and F1 value are compared by setting different thresholds, as shown in the figure 2,3,4,5.

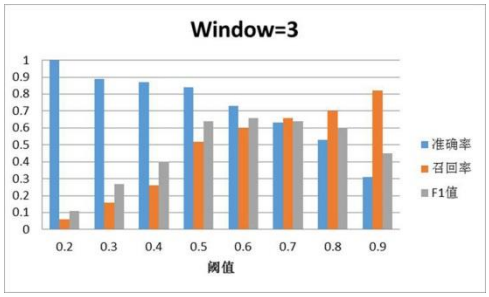

Figure 2: The window size is 3, the accuracy, recall and F1 value when the threshold is different.

When the window size is 3, as the threshold increases, the accuracy decreases gradually, and the recall rate gradually increases. The value of F1 increases first and then decreases. When the threshold is 0.6, the value of F1 is the maximum. At this time, the value of F1 is 0.66.

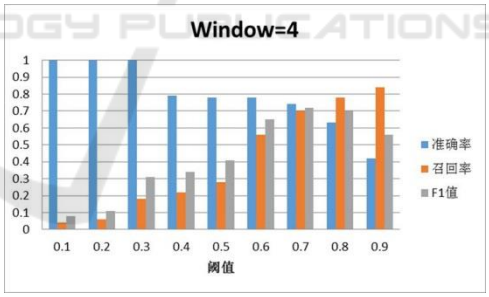

Figure 3: The window size is 4, the accuracy, recall and F1 value when the threshold is different.

When the window size is 4, as the threshold increases, the value of F1 increases first and then decreases. When the threshold is 0.7, the value of F1 is the maximum. At this time, the value of F1 is 0.72.

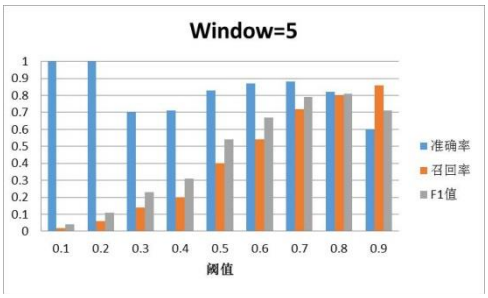

Figure 4: The window size is 5, the accuracy, recall and F1 value when the threshold is different.

When the window size is 5, as the threshold increases, the value of F1 increases first and then decreases. When the threshold is 0.8, the value of F1 is the maximum. At this time, the value of F1 is 0.81.

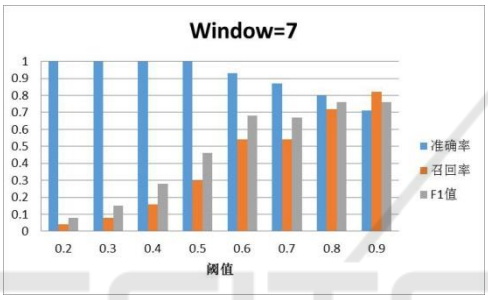

Figure 5: The window size is 7, the accuracy, recall and F1 value when the threshold is different.

When the window size is 7, as the threshold increases, the value of F1 increases first and then decreases. When the threshold is 0.9, the value of F1 is the maximum. At this time, the value of F1 is 0.76.

Comparing the maximum value of the F1 value when the window size is 3, 4, 5, and 7, it can be concluded that when the window size is 5 and the threshold is 0.8, the F1 value is the maximum value, and the experiment works best.

5 CONCLUSIONS

Scene division is the key component of the research of text-to-scene conversion. In this paper, a scene division method based on theme is proposed, and this method divides multi-scene texts by LDA model. The experiment results show that this method is correct and feasible. At present, the study of scene division in text-to-scene conversion is just some preliminary exploration. In the future, it is intended to use the objects in the scene to identify and divide multi-scene text, and to use the time and space elements in the scene as appropriate. The research of scene division is not only to identify the number of scenes in multi-scene text, but more importantly to recognize the scene structure of multiple scenes and the relationship between scenes. The reasoning of the scene relationship will be the focus of future work.

REFERENCES

Coyne, B., Sproat, R., 2001. WordsEye: An automatic text-to-scene conversion system. In: *28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001, Los Angeles August 12 – 17, 2001*, pp.487-496.

Seversky, L M., Yin, L., 2006. Real-time automatic 3D scene generation from natural language voice and text descriptions. In: *14th Annual ACM International Conference on Multimedia, MM 2006, Santa Barbara October 23 – 27, 2006*, pp.61-64.

Joshi, D., Wang, J Z., Li, J., 2004. The story picturing engine: Finding elite images to illustrate a story using mutual reinforcement. In: *MIR'04 - Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, New York October 15 – 16, 2004*, pp.119-126.

Giovanni, A., Mauro, D M., Fausto G., 1984. Natural language driven image generation. In: *10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics, Stanford, California July 02 - 06, 1984*, pp.495-500.

Clay, S R., Wilhelms, J., 1996. Put: language-based interactive manipulation of objects. *IEEE Computer Graphics and Applications*. 16, 31-39.

Chang, A X., Savva, M., Manning, C D., 2014. Interactive Learning of Spatial Knowledge for Text to 3D Scene Generation. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, January 2014*, pp.14-21.

Chang, A X., Savva, M., Manning, C D., 2014. Learning spatial knowledge for text to 3D scene generation. In: *2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar October 25 – 29, 2014*, pp.2028-2038.

Winograd T., 1971. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. *Computational Linguistics*. 20, 464.

Johansson, R., Williams, D., Berglund, A., Nugues, P., 2004. Carsim: a system to visualize written road accident reports as animated 3D scenes. In: *2nd Workshop on Text Meaning and Interpretation, Barcelona, Spain July 25 – 26, 2004*, pp.57-64.

Blei, D M., Ng, A Y., Jordan, M I., 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*. 3, 993-1022.